THE INTERNATIONAL JOURNAL OF HUMANITIES & SOCIAL STUDIES

Analysis of Psychometric Qualities of National Examinations Council (NECO) Mathematics Essay Test Using Generalized Partial Credit Model

Dr. Akobi, Thomas Ogbeche
Lecturer, Department of Science Education,
University of Nigeria, Nsukka, Nigeria
Ezugwu, Ifesinachi Jude
Assistant Lecturer, Department of Science Education,
University of Nigeria, Nsukka, Nigeria
Madu, Barnabas Chidi
Senior Lecturer, Department of Science Education,
University of Nigeria, Nsukka, Nigeria
Dr. Foluke Bosede Eze
Lecturer, Department of Science Education,
University of Nigeria, Nsukka, Nigeria
Stanley Terkuma Asongo
Postgraduate Students, Department of Science Education,
University of Nigeria, Nsukka, Nigeria

Abstract:
The main purpose of this study therefore was to analyze the psychometric qualities of NECO mathematics essay test using generalized partial credit model. The study adopted an instrumentation research design and was carried out in Benue state of Nigeria. The population of the study was 41,836 SS 3 students who registered for June/July NECO SSCE Examination in Benue State of Nigeria and a sample size of 650 students were used. The instrument for this study was an adopted NECO 2016 and 2017 June/July mathematics essay questions. The instrument for the study was administered to the respondents by the mathematics teachers in the sampled schools under the supervision of the researcher. Data was analysed using item parameter estimates of R-Package to answer the research questions, while t-test statistic was used to test the null hypotheses at 0.05 level of significant. The result of the study showed that NECO 2016 and 2017 mathematics essay questions were unidimensional in nature. NECO 2016 had lower step difficulties than NECO 2017. NECO mathematics essay test for 2016 discriminated between high and low ability students than the NECO mathematics essay test for 2017. Based on the findings of the study, it was recommended among others that test developers and examination bodies such as NECO should undergo rigorous processes of psychometric properties of essay test before and after administration of test instrument. This will help the test developers to determine the existence of these psychometric properties and the need to establish them in any given test. It was also recommended that Seminars and workshops on Item Response Theory should be organized by the examination bodies for Psychometricians, test developers and teachers involved in test development.

Keywords: *Mathematics, psychometric, assessment, test, item difficulty, item discrimination*

1. Introduction

Mathematics is considered as an indispensable tool needed for the transformation of technological development to reality. This is in agreement with Anastacio (2007) who described mathematics as the supporting knowledge of modern sciences. This means that scientific development depends on a deep understanding of mathematical concepts and procedures, hence a meaningful understanding of mathematics must occur. Mathematics is seen as a fundamental science that is necessary for understanding of most other fields in education. It is glaring that no other subject forms such a strong force among the various branches of science. Odusoro (2002) affirmed that the knowledge of science remains shallow without mathematics. It therefore means that, the position of mathematics in secondary school curriculum is very important for scientific development of any nation.

Mathematics is very important and useful in most fields of human endeavours. Its usefulness in science and technological activities as well as education and even humanities cannot be underestimated. It is one of the key subjects in both the primary and secondary school education system in Nigeria. It is one of the compulsory subjects that students

must offer in senior secondary school, not minding whether such students are in science, commercial, arts or social science classes. It is one of the essential subjects for students' advancement.

Despite the importance of mathematics towards achieving scientific and technological developments, it is sad to observe that students' achievement in the subject has remained very low especially in Senior Secondary School Certificate Examination. Studies such as Agashi (2014), Agwagah and Utibe, (2015), Galadima and Yushau (2007) and Uloko and Usman (2007) separately indicated that there is a poor and fluctuating state of students' academic achievement in mathematics. Evidence in literature has also shown that most senior secondary school students achieve poorly in many topics of the Mathematics curriculum (Ali, 2016). This is evidenced in the Chief Examiner's Reports of the National Examination Council (NECO) which consistently reported a poor and fluctuating achievement of students in mathematics between 2015 and 2018.

However, the conduct of the Senior School Certificate Examinations (SSCE) which had, hitherto, been the exclusive preserve of the West African Examinations Council (WAEC) was made an additional responsibility of the new examination body (i.e. National Examination Council (NECO)). NECO was to take exclusive charge of the conduct of the SSCE for school based candidates while WAEC was to take charge of the same examination for private candidates. The stakeholders in education express worry over the growing rate of failure in mathematics in senior secondary schools in recent times. The growing failure rate could be noticed in the yearly poor performance by students in the Senior School Certificate Examination (SSCE) in mathematics and this to some extent has affected the quality of education provided for the Nigerian citizens.

The quality of education in a given country can therefore predict the progress of that nation. Education therefore plays an important role in human development and is associated with an individual's well-being and opportunities for better living. It is for these reasons that Nigeria in her National Policy on Education adopted education as an instrument 'par excellence' for effecting national development as well as harnessing the potentials of the citizens (Federal Republic of Nigeria, [FRN] 2014). Education is therefore seen as a way of answering so many questions and solving so many problems confronting the nation. Considering the importance of education in national development, there is need to ensure its quality through valid assessment tools.

Assessment is a term that is commonly used in several areas of human endeavours. In education, assessment is an essential component of teaching. Nworgu (2015) defined assessment as 'a systematic process of gathering data from a variety of sources in order to understand, describe and improve learning'. Assessment is also used for the purposes associated with the placement of students, the award of qualifications and monitoring students' achievement and progress (Adonu, 2014). It is also defined as 'a systematic process for gathering data about student achievement' (Dhindsa, Omar, & Waldrip, 2007). These definitions imply that assessment refers to the full range of information gathered and synthesized by teachers about their students and instruction. It is a method for analyzing and evaluating student achievement or programme success. In essence, assessment is the process of gathering evidence to make inferences about how students are progressing toward specific goals.

Assessment can be formative or summative in nature. Formative assessment according to Nworgu (2015), is that assessment undertaken while a lesson, course or programme is still in progress or ongoing in order to collect relevant data and use the feedback to improve learning, course of programme. On the other hand, summative assessment (SA) or assessment of learning, according to Nworgu (2015) is that assessment carried out to determine what students have been able to learn at the end of a given lesson, unit, programme or period of schooling. The uses of summative assessment according to Nworgu (2015) include to measure what students have learnt at the end of a unit, promote students, ensure they have met required standards for certification on completion of a level or period of school, determine candidates who have met the required standards to enter certain occupations and to select students for entry into further education. Other uses of summative assessment have been identified to include sorting and ranking students, assigning grades to individual learners and evaluating programmes, curricula and schools (Hoover, 2013). Feedback from summative form of assessment is generated at the end of a programme when it is no longer feasible to use it to affect any modifications (Nworgu, 2015).

Furthermore, summative assessments are used to evaluate student learning at the end of a specific instructional period - typically at the end of a unit, course, semester, program, or school year. Summative assessments are scored and graded and are used to determine whether students have learned what they were expected to learn during the defined instructional period (Suskie, 2004). Assessment whether formative or summative may take the form of multiple choice or essay test.

Test refers to a structured situation comprising a set of items (i.e. questions or statements) with preferred responses, given to individuals or testees to determine the amount of the relevant trait or attribute they possess (Nworgu, 2015). Nworgu further stated that on the basis of an individual's response, his behaviour is quantified. A test is the most popularly used technique for obtaining information in the educational or school system. A test enables teachers to systematically obtain data for the purpose of making comparisons across individuals, classes, schools, districts or countries. Tests may take the form of multiple choice or essay test. In case of essay test which is the focus of this work, Nworgu (2015) stated that it is the type of test that permit testees to express their responses in their own words and in the ways they deem fit. This means that students have the responsibility of thinking out the answers to the questions. It is a free answer kind of test. It is used by teachers and examination bodies such as NECO to measure students' achievement. In this case, the students have the opportunity to write what they can based on their ability level and their knowledge of subject matter.

However, in spite of the numerous advantages of essay test, the achievement of students in public examinations on the essay test has not been encouraging especially in mathematics (Tata, Abba & Abdullahi, 2014). This has been a

source of concern to government, educators, parents and the general public considering the funds that are being committed to education by these stakeholders. Studies such as Moyinoluwa (2015) and Adonu (2014) have shown that the reasons for the low achievement by students could be attributed, among others, to the poor state of education in the country, low quality teaching staff, nature of the subjects, inadequate preparation of students for external examinations as well as cut in education budgets leading to shortages of facilities and equipment needed for effective teaching and learning. Several other factors observed to be responsible for students' mass failure in public examinations, are those related to home, society and parents' inability to inculcate discipline and learning habits in their children. Yet, other factors may be Government's failure to provide human and material resources to facilitate good teaching and learning (Adonu, 2014), while some factors may be teachers' inability to impact the necessary knowledge, skills and behaviour to students and also the students themselves may be blamed for refusal to learn, while majority of others may be blamed on external examination bodies such as West African Examinations Council (WAEC) and National Examinations Council (NECO) for their failure to measure adequately the knowledge, skills and behaviour learned by students.

In any case, since the senior secondary school certificate examinations are set, conducted, scored and graded by bodies external to the schools, it is possible that some of the factors that account for poor performance could be related to the external examination bodies. However, other factors that are responsible for students' low achievement in public examination which bother around the nature of mathematics tests items (that is, its psychometric properties such as reliability, item difficulty, discrimination, distractor indices and test information function among others) are often neglected by researchers. Psychometric analysis of a test is therefore crucial.

The psychometric analysis of a test, according to Adonu (2014), is the science of measuring latent traits or constructs in subjects of interest which involve analyses of constituents of a test such as validity, reliability, difficulty index, discrimination index and distractor index. The psychometric analysis of mathematics essay test therefore, implies analyzing the dimensionality, difficulty and discrimination indices, reliability and validity indices and the test information function. Psychometric tests are standard and scientific methods used to measure individuals' mental capabilities and behavioural style. The term psychometrics refers to the design and interpretation of tests that measure psychological values such as aptitude, ability, personality, memory and intelligence. This is because the relevance of an essay test largely depends on these psychometric properties of test items.

The dimensionality of a test, as defined by Lord and Novick (1968), is the total number of abilities required to satisfy the assumption of local independence. A set of items is called locally independent if, for fixed values of the latent traits, the item responses are statistically independent. For a weaker form of local independence to hold, it is sufficient that for fixed values of the traits the item responses are uncorrelated (McDonald, 1981). Many educational and psychological tests are inherently multidimensional, meaning the tests measure two or more constructs or dimensions. A construct is a theoretical representation of the underlying trait, concept, attribute, process, and/or structure that the test is designed to measure (Messick, 1989). The items on a factorially simple test measure one underlying dimension (McDonald, 1999). For example, one might believe that a mathematics test is measuring one identifiable construct algebra. However, the dimensional structure of most real testing data is much more complex. For example, one might suspect a mathematics test is measuring algebra and geometry. In this case, a subset of test items with algebra content might be considered a measure of the first dimension, whereas the remaining items with geometry content might be considered a measure of the second dimension. These items may have varied level of difficulty.

Item difficulty is also known as item facility or easiness. It is defined as the index that describes the level of the difficulty of a test item (Harbour-Peters, 1999). According to Nworgu (2015), the difficulty index provides answer to the question such as; how hard is the item? Perhaps, this is why it is referred to as difficulty or easiness index. An ideal item should have a difficulty index of 0.5, it could range from 0.3 to 0.7 (Nworgu, 2015). The difficulty index (or step difficulty with reference to essay test) of a test item that is reported for a particular test administered to a particular group is a function of the skills required by the questions and the skills achieved by those attempting the test. Habor-Peter (1999) relates the item difficulty or facility to the proportion of students answering each item correctly. It helps in ensuring that items that are suitable are included in the final version of the test. The difficulty index determines whether the items are difficult or easy.

The discrimination index hence reflects how each item distinguishes between candidates with the required knowledge and skill and those lacking such knowledge and skill. This is in line with Harbor Peters (1999) who viewed discrimination index as the measure of the extent to which a test item discriminates between high ability and low ability students who got the item right divided by the number in either group. Choosing items with an acceptable discrimination index will tend to provide new version of test with greater homogeneity (Adonu, 2014). Hence, there is a relationship between mean discrimination index and the reliabilities of the test. The higher the mean discrimination index, the higher the reliability coefficient of the test.

These psychometric properties of a test can be measured or determined using either Classical Test Theory (CTT) or Item Response Theory (IRT). In CTT, the total test score in terms of number of correct responses to the items, has a central role both for item analysis and for student evaluation. One of the main drawbacks of CTT is that the evaluation of student performance is strongly influenced by the sample analyzed. On the premise of having weak theoretical assumptions, the CTT has been seen as not being as precise as item response theory (IRT) for ensuring objectivity in psychometric analyses (Adonu, 2014).

The CTT has many limitations, one of which includes weak theoretical assumptions that cast doubts when psychometric properties of tests are obtained. There is therefore, the need to change the method of analysis of psychometric properties of tests from CTT to a theory that will further reduce the shortcomings of CTT model. In

particular, there is the need to study the psychometric properties of mathematics essay test in NECO as many have been done in multiple choice test. Despite the importance of mathematics in National development, the analyses of psychometric properties of mathematics essay test with particular reference to the use of Generalized Partial Credit Model (GPCM) has remained scarce in recent times. Various studies in the past such as Obinne (2011), Obinne (2008), Nworgu (1985), Agwagah (1985), Obioma (1986) went variously into psychometric analyses of questions that are dichotomously scored. It is therefore important to test the psychometric properties of NECO mathematics essay test using the generalized partial credit model of Item Response Theory (IRT).

The generalized partial credit model (GPCM) was formulated by Muraki (1992) based on Masters' (1982) partial credit model (PCM) by relaxing the assumption of uniform discriminating power of test items. In Masters' PCM, the discrimination power is assumed to be common for all items. This model is a member of the Rasch family of item response models. The GPCM not only can attain some of the objectives that the Rasch model achieves but also can provide more information about the characteristics of test items than does the Rasch model. The generalized partial credit model (GPCM) is an IRT model developed to analyze partial credit data, where responses are scored 0, 1, k, where k is the highest score category for the item. The Generalized Partial Credit Model is a generalization of the partial credit model that allows discrimination parameter to vary among the items. According to Tang (1996), the major difference between the partial credit model and the Generalized Partial Credit Model (GPCM) is that the partial credit model assumes that the item discrimination is a constant for all the items in a test. The generalized partial credit model on the other hand assumes that item discrimination can be different across items.

The difference between these two models is similar to the difference between Rasch model and the two parameter logistic model (2PLM) in the dichotomous case. Tang (1996) noted the following while discussing parameter interpretation in GPCM that both the partial credit model and GPCM assume that each of the two adjacent categories (k and k-1) in a polytomously scored item can be seen as dichotomous categories and therefore, the likelihood of a person with certain ability level reaching the score category of k rather than k-1 can be described by a dichotomous IRT model. For polytomously scored item that have *m* score categories, based on GPCM, the items have one item discrimination parameter, one location parameter and a set of m-1 threshold parameters. That is to say that in GPCM we have only one difficulty index (*b)*, one discrimination index (*a)*. The item discrimination parameter describes how well the item can distinguish between individual of different ability levels. The location parameter indicates the item difficulty. All these have implications on students' achievement in mathematics especially the essay test, hence the need to analyze the psychometric qualities of NECO mathematics essay test using Generalized Partial Credit Model.

1.1. Research Questions

The following research questions were posed to guide the study

- What is the dimensionality of NECO mathematics Essay Test for 2016 and 2017?
- What are the step difficulty indices of NECO mathematics essay for 2016 and 2017?
- What are the discrimination indices of NECO mathematics essay test for 2016 and 2017?

2. Literature Review

2.1. Studies on Dimensionality of Test

Adonu (2014) carried out a study to analyse the psychometric qualities of practical physics questions of West African Examination Council and National Examinations Council using the Partial Credit Model (PCM). The objectives of the study were specifically to evaluate the Standard Error of Measurement (SEM), the fit statistics and item difficulty estimates of WAEC and NECO practical physics items and also to test for significant difference of NECO, WAEC and NECO-WAEC psychometric qualities in various years. The design of the study was instrumentation research design and the area of the study was Enugu State of Nigeria. The population of the study was all SS III physics students of 2012/2013 academic session in Enugu State. A sample of 668 physics students was drawn through multi stage sampling procedure. The instrument for the study consisted of four different tests viz; two practical physics questions of NECO 2011 and 2012 (PPQN, 1 and 2) and two practical physics questions of WAEC 2011 and 2012 (PPQW, 1 and 2). The research questions were answered using the descriptive statistics of Winstep software maximum likelihood ratio. The hypotheses were tested at 0.05 level of significance using independent sample t-test statistic, and the chi square goodness of fit test using WINSTEP PCM analysis and SPSS. The major findings of the study indicated that: The standard error of measurement (SEM) of items of WAEC and NECO practical physics in 2011 and 2012 were very low- below 0.18 for all items. The fit statistic indicated that nearly all the items of both examinations NECO and WAEC were valid and thus sufficiently demonstrated unidimensionality; the item difficulty estimates (b) for both examinations for the two years studied showed that all the items have difficulty estimates that range between 0.3 to 0.7 which show that their difficulty are moderate for all items. All the four different tests that constituted the instrument had very high proportion of their item fit to PCM with all the four parts having 0.92 proportion of fit.

Grazielle (2017) carried out a study to evaluate the psychometric properties of an instrument to assess comprehensiveness of care from dentists using a combination of classical test theory and item response theory in Brazil. A 46-item instrument was developed and tested by a panel of experts, followed by a pilot test and administration to 187 primary care dentists in a large Brazilian city. The 46 items were evaluated using the following criteria: acceptability, internal consistency, temporal stability, inter-item correlation, and tetrachoric correlation. This evaluation led to a shortened version consisting of 11 items that met all the criteria previously described. The temporal stability was

measured using Cohen's kappa, and all 11 items presented values greater than 0.5. The Cronbach's alpha value was 0.72. None of the 11 items had missing data on the distribution of responses, and the model considering the discrimination as varying fit the data better than the model considering discrimination as a constant parameter ($p<0.001$). The result showed that the items of the test were unidimensional in nature. Item characteristic curves showed that 54.5% of items could be considered difficult, i.e., only dentists with a good understanding of comprehensiveness responded favorably.

2.2. Studies on Item Difficulty and Discrimination Indices of a Test

Nworgu and Agah (2012) carried out a study on the application of three parameter logistic model (3PLM) in the calibration of a mathematics achievement test. The purpose of the study was to use 3PLM of item response theory in the calibration of a mathematics achievement test. Three research question and three hypotheses guided the study. The study used 1514 SS III students from Rivers and Cross river states of Nigeria as sample. The instrument for data collection was a 40 items multiple choice test developed by the researchers. The data analysis was carried out using BILOG-MG an IRT computer software that estimated the item parameter and their corresponding standard error of measurement. The chi square goodness of fit was used to determine the goodness of fit of the items of the instrument to three parameter logistic model. The study also generated item characteristics curve to determine if the items in the tests are good enough for the assessment of the students' ability. The result showed the empirical reliability coefficient of 0.79. The item parameter indices obtained indicated that the discrimination parameter (a) ranges from 0.29 to 2.05; item difficulty form -0.40 to 1.79; the probability of guessing in the test correctly ranged from 0.02 to 0.50 for all the ability levels. While the study by Nworgu and Agah used the 3PLM for multiple choice items, the present study will use generalize partial credit model which is an extension of 2PLM for essay test.

Obinne (2008) carried out a study to compare the psychometric properties of WAEC and NECO Biology examinations under item response theory (IRT). The study intended to compare the reliability, validity, difficulty index of multiple choice items of biology examinations conducted by WAEC and NECO using the item response theory (IRT). Fourteen research questions and seven null hypotheses guided the study. The sample of the study was 1800 SS III students from 36 secondary schools in urban and rural areas of Benue State. Multistage stage sampling procedure was used to select the sample for the study. WAEC and NECO biology examination questions from the year 2000 – 2002 were the instruments used for data collection. The research questions were answered using maximum likelihood estimation technique of BI-LOG MG computer programme according to IRT procedure. The SPSS was used to test the hypotheses at 0.05 level of significance. The results of the study among others showed that biology examination items from WAEC and NECO were equally reliable and valid; and that biology items of NECO examination were more difficult than those of WAEC of the same year. The study concluded that NECO questions were really not easy to pass. The study also discovered that WAEC items were more prone to guessing than those of NECO items.

Obinne (2011) carried out a study to determine the psychometric properties of two major examinations in Nigeria – SSCE by WAEC and SSCE by NECO. The aim of the study was to compare the standard error of measurement (SEM) of biology examination conducted between the year 2000 to 2002 by WAEC and NECO using one parameter logistic model. The study used Instrumentation research design. The area of study was Benue State of Nigeria. The population of the study was all 33,541 year three (SS III) students who registered for May/June 2006 biology examination in WAEC and NECO in Benue State. The sample of the study was 1800 students. The researcher used the objective Biology questions for the years 2000-2002 as instrument for data collection. Maximum likelihood estimation techniques of the BILOG MG were used for data analysis. The result of the study among others indicated significant difference ($p<0.05$) in the SEM of NECO and WAEC Biology in the years understudied. The result also showed that Biology examination conducted by NECO had smaller SEM than those of WAEC and noted that NECO Biology has higher reliability than that of WAEC.

Ugodulunwa and Muttsapha (2011) used differential Item Functioning (DIF) analysis for studying the improvement of quality in State Wide examination in Nigeria. The problem of threat to validity of JSCE as a result of low correlations between results of JSCE and SSCE in Mathematics that is replete in literature necessitated this study. Cluster sampling was used to select eleven local government areas and 77% of all the scripts used for JSCE examination in mathematics in the years 2007 and 2008. A total of 27038 scripts formed the sample for the study. Six hypotheses guided this study. The data analysis was done using Scheuneman modified Chi-square statistics to identity the presence or otherwise of DIF in the mathematics items that were dichotomously scored. The findings of the study were that the examination items contained items that differentially functioned for candidates described by gender, school type and school location. The result of the study also showed that 2007 mathematics questions were more difficult than 2008 questions. The study recommended that to ensure quality in a state wide examination such as JSCE, DIF analysis should form part of the test process and in fact in other nation-wide examinations in Nigeria.

Akindele (2004) worked on development of prototype of items for selection tests into universities in Nigeria. Using computer programme he randomly generated a sample of a thousand students who entered for 1998 university entry examination-made up of 626 males and 374 females-in English Language. The data analysis was done using SPSS and BILOG MG software. The SPSS accomplished the classical item statistics, while the BILOG MG software was used to calibrate the test to determine the item parameters and ability estimates. Test of hypothesis revealed significant differences in the item parameter estimates of test items using IRT and CTT; but the scaled scores for the three subparts of the test (grammar, lexis and structure, comprehension) did not show any significant difference in the mean and standard deviations as computed using CTT and IRT procedures. Three different ability estimation procedure used in the study did not reveal any significant difference in estimated abilities of the students. Gender was noted in the study as a moderating variable in the student academic performance as it established differential item functioning. The values of item statistics a,

b and c as estimated using 1, 2 and 3 parameter logistic models of IRT showed significant difference. The item developed and stored in this study's item bank was calibrated with 3-PL model because the study deemed it (3-PLm) to be more robust given the sample size and the length of the test. The difference between Akindele's work and the present study is that Akindele used 3PLM because of the guessing parameter while the present study used 2PLM because of GPCL.

3. Methodology

The study adopted an instrumentation research design. According to international Centre for Educational Evaluation (ICEE) (1982), instrumentation research is a study aimed at introducing new or modified content, procedure, technology or instruments of educational practice. The study was conducted in Benue State, Nigeria. Benue State is one of the middle belt States in Nigeria. The population of the study was 45,934 SS 3 students. This comprised all senior secondary school three mathematics students (SSIII) who registered for June/July, 2019 senior secondary certificate examination (internal examination) of NECO in the State owned secondary schools (public) in the three educational zones of Benue State. The sample size for the study is 650 SS3 students.

The instrument for this study was adopted NECO 2016 and 2017 June/July mathematics essay questions. The adopted instrument is made up of 12 essay questions each. According to the examination's instruction, students are expected to attempt 10 questions out of 12. But for this study, students were expected to attempt all questions. The items of the instruments were re-typed and numbered serially to conform to the Item Response Theory. After re-typing, NECO 2016 had 26 items and NECO 2017 had 29 items. NECO marking scheme for essay test was used by the researcher and the research assistants who were mathematics teachers in the sample schools to mark and score the test.

The instrument for the study was administered to the respondents by the mathematics teachers in the sampled schools under the supervision of the researcher. The researcher ensured adequate supervision to check examination malpractice. The invigilators of the examination also ensured strict compliance of respondents to instruction. Similar conditions of administration by NECO was adopted. The study was carried out in second half of third term when the SS3 students were ready for NECO examinations. NECO 2016 was first administered to the students, while the NECO 2017 was administered to the students two weeks after the administration of the first test. The scores were used for data analysis. The marking guide for the questions was the marking scheme provided by NECO.

The research questions two and three were answered using item response theory descriptive statistics estimation procedure such as mean, Standard Error of Measurement (SEM), item parameters estimate, to test for dimensionality, Stout's (1987) test of essential unidimensionality was used to answer research question one. According to Stout, the dimensionality of a test data is judged based on the following:

Strong multidimensionality	DETECT > 1.00
Moderate multidimensionality	.40 < DETECT < 1.00
Weak multidimensionality	.20 < DETECT < .40
Essential unidimensionality	DETECT < .20
Maximum value under simple structure	ASSI=1, RATIO=1
Essential deviation from unidimensionality	ASSI > .25 RATIO > .36
Essential unidimensionality	ASSI < .25 RATIO < .36.

In IRT, the higher the difficulty index, the more difficult the item and the lower the difficulty index, the easier the item and it ranges between -3 to +3. This range of value was used to judge the difficulty index of the items for research question one. Also, for item discrimination, a higher value indicates that the item discriminates (differentiates) between high and low ability examinees, it also ranges between -3 to +3. This range of values were also used to dictate items that discriminate between low and high ability students for research question three.

4. Results

4.1. Research Question 1

What is the nature of dimensionality of NECO mathematics Essay Test for 2016 and 2017?

The result of the dimensionality assessment of the 2016 and 2017 NECO Mathematics essay tests are presented in Table 1.

	NECO 2016		**NECO 2017**	
	Unweighted	**Weighted**	**Unweighted**	**Weighted**
DETECT	-1.9802927	-1.9802927	-0.0204	-0.0204
ASSI	-0.1015385	-0.1015385	0.0000	0.0000
RATIO	-0.1744425	-0.1744425	-0.0018	-0.0018

Table 1: Unidimensionality Assessment of 2016 and 2017 NECO Mathematic Essay Test
Note: DETECT: Dimensionality Evaluation to Enumerate Contributing Traits, ASSI: Application Software Solutions Inc

Table 1 shows that the 2016 NECO Mathematics test was essentially unidimensional (maximum DETECT value = -1.9802927 (< .20), ASSI = -0.1015385 (< 0.25) and RATIO = -0.1744425 (< 0.36)). Therefore, the assumption of unidimensionality was met. This result showed that one dominant dimension accounted for the variation observed in students' responses to the Mathematics essay test items. The implication of the result is that the 2016 NECO Mathematics essay test was unidimensional. Table 1 also shows that the 2017 NECO Mathematics test was essentially unidimensional

(maximum DETECT value = -0.0204 (< .20), ASSI = 0.0000 (<0.25) and RATIO = -0.0018 (< 0.36)). Therefore, the assumption of unidimensionality was also met implying that one dominant dimension also accounted for the variation observed in students' responses to the NECO Mathematics essay test items in 2017. This also implies that the 2017 NECO Mathematics essay test was unidimensional. The overall finding therefore shows that the NECO 2016 and 2017 essay questions were unidimensional in nature and this indicated the measure of a particular trait which maybe cognitive ability of the students.

4.2. Research Question 2

What are the step difficulty indices of NECO mathematics essay for 2016 and 2017?

NECO 2016					**NECO 2017**					
Items	B1	B2	B3	B4	Items	B1	B2	B3	B4	B5
1	2.31	-2.92	0.71	--	1	2.19	-2.67	2.21	0.27	0.18
2	-1.55	0.18	2.25	--	2	2.79	1.54	2.83	-2.50	-0.43
3	2.01	1.44	-2.59	--	3	1.60	-2.63	2.92	--	--
4	2.68	-0.08	-1.23	0.12	4	0.72	-1.49	-1.50	2.20	--
5	-2.26	1.41	2.73	--	5	-1.03	-0.85	--	--	--
6	-1.83	0.14	0.44	1.92	6	-0.73	1.19	1.62	--	--
7	2.09	-1.67	1.30	--	7	-0.76	-1.87	1.78	-2.27	--
8	2.54	2.37	-1.80	-0.72	8	1.95	-0.51	-0.33	--	--
9	-2.05	1.73	1.02	--	9	-0.87	1.93	2.36	--	--
10	-2.82	0.72	0.22	--	10	0.88	-2.37	-1.22	--	--
11	-1.09	0.71	1.69	--	11	2.33	-2.45	-1.98	--	--
12	-1.04	-0.38	1.49	0.85	12	-2.05	-2.52	1.62	--	--
13	-2.08	1.86	2.06	--	13	1.71	1.54	-2.68	-2.74	-0.14
14	-1.81	-0.27	1.54	2.03	14	0.09	0.89	--	--	--
15	-1.85	-0.95	2.67	0.68	15	1.75	-1.90	--	--	--
16	-1.72	-0.47	2.04	1.08	16	1.46	0.31	2.44	--	--
17	-2.83	-0.52	1.27	2.14	17	2.49	--	--	--	--
18	2.86	-2.27	1.71	0.04	18	1.27	2.26	1.83	-2.87	--
19	-0.27	1.40	1.99	--	19	2.08	-1.74	2.24	--	--
20	0.49	1.34	2.74	--	20	2.86	1.18	--	--	--
21	-0.22	-0.20	2.21	2.60	21	0.08	-0.33	--	--	--
22	1.54	1.76	2.75	0.32	22	2.60	-1.79	--	--	--
23	0.62	1.16	0.15	--	23	2.29	-2.71	--	--	--
24	-2.77	1.38	1.81	--	24	2.25	--	--	--	--
25	-0.39	1.39	1.66	--	25	2.45	--	--	--	--
26	1.56	2.17	1.80	1.31	26	2.91	--	--	--	--
					27	2.86	-0.71	2.52	--	--
					28	2.19	-1.56	--	--	--
					29	1.44	-2.03	--	--	--

Table 2: Step Difficulty ('B' Parameter Estimates) of the 2016 And 2017 NECO Mathematics Essay Test Based On Generalized Partial Credit Model.
Note: -- = Steps Not Attempted

Result in Table 2 shows the step difficulty indices of NECO mathematics essay test for 2016 and 2017. In IRT, the higher the difficulty index, the more difficult the item and the lower the difficulty index, the easier the item. Result shows that for item 1 in NECO 2016, b1 = 2.31, b2 = -2.92 and b3 = 0.71, this shows that step1 was difficult for the examinees, step2 was easy and step3 was difficult, which means that item 1 was difficult. Step difficulty indices for item 2 ranges from b1 = -1.55 to b3 = 2.25, it also shows that item 2 was difficult. Item 3 has b1 = 2.01, b2 = 1.44 and b3 = -2.59 which shows that step 3 was easy for the examinees while step 1 and 2 were difficult. In item 4, only step 1 was difficult while steps 2,3 and 4 were easy. In item 5, step difficulties indices were; b1 = -2.26, b2 = 1.41 and b3 = 2.73. This means that step 1 was easy while steps 2 and 3 were difficult. For item 6, b1 = -1.83, b2 = 0.14, b3 = 0.44 and b4 = 1.92, the step difficulties increase with increasing steps involved in solving it. In item 8, steps 1 and 2 were difficult while steps 3 and 4 were easy. From Items 9 – 17, steps 1 and 2 were easy for all the items while step 3 and 4 were somewhat difficult for the examinees with exception of step 3 for item 10 which was easy. The overall finding of the study shows that for NECO 2016, step 1 was easy for majority of the examinees while the step difficulty indices increase with increasing steps in solving a particular item. The overall finding of the study shows that NECO 2016 was difficult given the performance of students on each item.

Result on Table 2 also shows that in NECO 2017, item 12 was the easiest at step 1 (b1 = -2.05) while item 26 was the most difficult item in step 1 (b1 = 2.91). In step 2, item 1 was the easiest (b2 = -2.67) while item 18 was hardest (b2 = 2.26). In step 3, item 3 was very difficult (b3 = 2.92) while item 13 was very easy (b3 = -2.68). In step 4, item 18 was easy (b4 = -2.87) while item 4 was difficult (b4 = 2.20). In step 5, the three items attempted by the examinees were easy. In

comparing the step difficulties for NECO 2016 and NECO 2017, result shows that NECO 2016 was relatively easier than the NECO 2017, this is because majority of the examinees attempted the various steps involved in solving the test items in NECO 2016 than in NECO 2017. This therefore suggests that NECO 2016 has lower step difficulties than NECO 2017.

4.3. Research Question 3

What are the discrimination indices of NECO mathematics essay test for 2016 and 2017?

NECO 2016		NECO 2017	
Item	**Discrimination Indices (a parameter)**	**Item**	**Discrimination Indices (a parameter)**
1	0.76	1	0.41
2	0.80	2	-0.18
3	-0.08	3	0.11
4	-1.15	4	0.49
5	0.08	5	-0.69
6	1.08	6	1.35
7	0.02	7	-0.33
8	-0.09	8	-0.49
9	0.65	9	0.31
10	0.77	10	-0.17
11	1.72	11	-0.12
12	0.50	12	0.14
13	0.60	13	-0.05
14	0.55	14	2.85
15	1.07	15	0.00
16	1.56	16	0.60
17	0.62	17	-0.79
18	0.25	18	-0.03
19	1.32	19	0.09
20	0.64	20	0.28
21	0.94	21	-1.09
22	0.23	22	-0.43
23	0.92	23	-0.35
24	0.17	24	0.68
25	0.33	25	0.58
26	0.56	26	-1.26
		27	0.39
		28	-0.71
		29	-0.01
	Mean = 0.57		Mean 0.05

Table 3: Discrimination Indices ('A' Parameter Estimates) of the 2016 And 2017 NECO Mathematics Essay Test Based on Generalized Partial Credit Model

The result of the study as presented on Table 3 shows the discrimination indices of NECO mathematics essay test for 2016 and 2017. For item discrimination, a higher value indicates that the item discriminates (differentiates) between high and low ability examinees. Result on Table 3 shows that for NECO 2016, the highest discrimination was 1.72 for item 11 (Logarithm) and the lowest discrimination was -1.15 for item 4 (geometry). When an item has a high discrimination, high-ability examinees have a much higher probability of answering it correctly than low ability examinees. According to DeMars (2010), a discrimination index of 0.20 and above is recommended, therefore, items 3, 4, 5, 7, 8 and 24 had discrimination indices less than 0.20. This implies that those items did not discriminate between high and low ability students. However, items 1, 2, 6, 9-23 and 25-26 had discrimination indices higher than 0.20, which means that the items discriminate between high and low ability students.

Also, in NECO 2017, the highest discrimination was 2.85 for item 14 (Geometric progression), and the lowest discrimination was -1.26 for item 26 (Graph). Result also shows that for NECO 2017, items 1, 4, 6, 9, 14, 16, 20, 24, 25 and 27 had discrimination indices greater than 0.20 while items 2, 3, 5, 7, 8, 10 – 13, 15 – 19, 21 - 23, 26 and 28 – 29 had discrimination indices less than 0.20. These imply that for NECO 2017, majority of the items did not discriminate between high and low ability students.

4.4. Hypothesis

- $H0_1$: There is no significant difference between the discrimination index of 2016 and 2017 NECO Mathematics tests items

NECO	N	Mean	SD	t	df	p-value	Dec.
2016	26	0.57	0.59	2.753	53	0.01	S
2017	29	0.05	0.78				

Table 4: T-Test Analysis of the Difference between the Discrimination Index of 2016 and 2017 NECO Mathematics Tests Items
S = Significant

The result of the study as presented in Table 4 shows the t-test analysis of the significant difference between the discrimination index of 2016 ($\bar{x}$ = 0.57, SD = 0.59) and 2017 ($\bar{x}$ = 0.05, SD = 0.78) NECO Mathematics tests items. Result shows that a t-value of t(53) = 2.753 was obtained with a probability value of 0.01. Since the p-value is less than 0.05 set as level of significance, the null hypothesis is rejected and inference drawn is that there is a significant difference between the discrimination index of 2016 and 2017 NECO Mathematics tests items with NECO mathematics essay test items having higher discrimination index than NECO 2017 items.

5. Discussion of Findings

The result of the study shows that the 2016 and 2017 NECO Mathematics tests were essentially unidimensional. This is because the maximum dimensionality Evaluation to Enumerate Contributing Traits (DETECT) values were less than 0.20. Therefore, the assumption of unidimensionality was met. This implies that one dominant dimension (like cognitive ability) or trait accounted for the variation observed in students' responses to the Mathematics essay test items for both 2016 and 2017 NECO examination. The implication of the result is that both 2016 and 2017 NECO Mathematics essay test were essentially unidimensional in nature. The findings of the study were consistent with Adonu (2014) who carried out a study to analyze the psychometric qualities of practical physics questions of West African Examinations Council and National Examinations Council using the Partial Credit Model (PCM) and found among other things that nearly all the items of both examinations (NECO and WAEC) were valid and thus sufficiently demonstrated unidimensionality. The finding of the study is also in agreement with Grazielle (2017) who carried out a study to evaluate the psychometric properties of an instrument to assess comprehensiveness of care from dentists using a combination of classical test theory and item response theory and found among other things that the items of the test were essentially unidimensional in nature. The results of this study therefore show that one dominant dimension (trait) accounted for the variation observed in students' responses to the Mathematics essay test items for both 2016 and 2017 NECO examination.

The results of the study show that both NECO mathematics essay test for 2016 and 2017 were difficult. This is because the step difficulty indices for 2016 and 2017 were high. It is important to note that in Item Response Theory (IRT), the higher the difficulty index, the more difficult the item and the lower the difficulty index, the easier the item. Results therefore show that for NECO 2016, the step difficulty indices were high which result in poor performance by the students. Result also shows that the step difficulty indices for NECO 2017 were also high which also result in poor performance by the study. It is also important to note here that for NECO 2016, step 1 was easy for majority of the examinees while the step difficulty indices increased with increasing steps in solving a particular item. In comparing the step difficulties for NECO 2016 and NECO 2017, result shows that NECO 2016 was relatively easier than the NECO 2017, this is because majority of the examinees attempted the various steps involved in solving the test items in NECO 2016 than in NECO 2017. This therefore suggests that NECO 2016 has lower step difficulties than NECO 2017. The overall finding of the study shows that mathematics essay test for NECO 2016 and 2017 were difficult. The result of the study is consistent with Obinne (2008) who carried out a study to compare the psychometric properties of WAEC and NECO Biology examinations under item response theory (IRT) and found among other things that biology items of NECO examination were more difficult than those of WAEC of the same year. Obinne also found that NECO questions were really not easy to pass. However, the finding of the study is not consistent with Adonu (2014) who carried out a study to analyse the psychometric qualities of practical physics questions of West African Examinations Council and National Examinations Council using the Partial Credit Model (PCM) and found that the item difficulty estimates (b) for both examinations for the two years ranged between -1.53 to +1.94 which showed that their difficulty indices were moderate for all items. The overall finding of this study therefore shows that NECO mathematics essay tests for the year 2016 and 2017 were very difficult with high step difficulty indices. This may also contribute to students' poor performance in NECO mathematics examination in general.

Research question three was aimed at finding the discrimination index of mathematics essay test for NECO 2016 and 2017. For item discrimination, a higher value indicates that the item discriminates between high and low proficiency examinees better. Result shows that for NECO 2016, the highest discrimination was 1.72 for item 11, and the lowest discrimination was -1.15 for item 4. Also, in NECO 2017, the highest discrimination was 2.85 for item 14, and the lowest discrimination was -1.26 for item 26. When an item has a high discrimination, high-proficiency examinees have a much higher probability of answering it correctly than low proficiency examinees. This assertion is in agreement with Nworgu (2015) who stated that a discrimination index of 0.20 and above is recommended. Result shows that the items of NECO 2016 discriminated between high and low examinees than the items of NECO 2017. This may be as a result of the difficult nature of NECO 2017. These imply that for NECO 2017, majority of the items did not discriminate between high and low ability students. To test for the significant different between the discrimination indices of NECO 2016 and 2017, an independent sample t-test statistic was used and the result showed that there was a significant difference between the discrimination indices of 2016 and 2017 NECO Mathematics test items. The finding of the study is somewhat consistent with Nworgu and Agah (2012) who carried out a study on the application of three parameters logistic model (3PLM) in the calibration of a mathematics achievement test and found among others that the discrimination parameter (a-parameter)

ranges from 0.29 to 2.05 which indicated high discrimination between the high and low ability examinees. However, the finding of this study showed that mathematics essay test for NECO 2016 discriminated between low and high ability students than NECO 2017.

6. Conclusion and Recommendations

6.1. Conclusion

Based on the findings of the study, the following conclusions are drawn.

NECO 2016 and 2017 essay questions were unidimensional in nature. NECO 2016 has lower step difficulties than NECO 2017. NECO mathematics essay test for 2016 discriminated between high and low ability students than the NECO mathematics essay test for 2017.

6.2. Recommendations

The following recommendations have been made based on the findings of this study.

- Test developers and examination body such as NECO should undergo rigorous processes of item analysis of essay test before and after administration of test instrument to determine their psychometric properties. This will help the test developers to determine the existence of these psychometric properties and the need to establish them in any given test in order to improve test quality and students' performance.
- Seminars and workshops on Item Response Theory should be organized by the examination bodies for Psychometricians, test developers, teachers involved in test development to help them understand the need for item analysis before and after the examination.
- The government, ministries of education and the examination bodies should purchase the various IRT analytical software and sponsor the training of test developers and psychometricians to learn how to use the software using IRT framework. This will make the interpretation of the results and usage of the framework simple and easy.
- Teachers at various levels of education in Nigeria should be train through workshops and seminars on the usage of IRT for psychometric analysis of their examinations. In this way, the quality of test items will get more refined and measurement problems associated with test construction will be solved.

7. References

i. Adonu, I. I. (2014). *Psychometric analysis of WAEC and NECO practical physics tests using partial credit model. (*Unpublished Ph.D Thesis), Department of Science Education, University of Nigeria, Nsukka.
ii. Agashi, P. P. (2014). *Effect of advance organizer and content attainment models on the achievement and retention of Pre-NCE students in geometry*. (Unpublished Ph.D Thesis), Department of Science Education, University of Nigeria, Nsukka.
iii. Agwagah, U.V. (1985). *The development and preliminary validation of an achievement test in mathematics methods*. M.Sc. Dissertation. University of Nigeria.
iv. Agwagah, U. N. V. & Utibe, U. V. (2015). A decade of candidates' performance in NECO-SSCE mathematics in Nigeria. *Journal of Education and Practice, 6*(25), 25 - 30.
v. Akinlele, P.B. (2004). *The development of item bank for selection test into Nigeria Universities*. (Unpublished Ph.D Thesis) Institute of Education. University of Ibadan.
vi. Anastacio, F. (2007). An Electronic-Workbook. *29th ASEE/IEEE Frontiers in Education Conference 12* 5-20
vii. Dhindsa, H., Omar, K., & Waldrip, B. (2007. Upper secondary bruneian science students' perceptions of assessment. *International Journal of Science Education, 29*(10), 1281-1280.
viii. Federal Republic of Nigeria (2014). *National Policy on Education.* Abuja: National Educational Research Council NERC Press.
ix. Galadima, I. & Yushua, M. A. (2007). Investigation into performance of senior secondary school students in Sokoto State. *Abacus, 32*(1), 24 - 33.
x. Grazielle, M. (2017). Evaluating psychometric properties of an instrument addressing comprehensiveness of care among dentists. *Brazilian Dental Journal, 28,* 638-646.
xi. Harbor-Peters, V.F.A. (1999). *Noteworthy points on measurement and evaluation.* Enugu: Snaap Press Ltd.
xii. Hoover, R. N. (2013) Teachers' instructional use of summative student assessment data. *Applied Measurement in Education, 26(*3), 219-231
xiii. Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Welsley Publishing Company.
xiv. McDonald, R. P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Erlbaum.
xv. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York, NY: American Council on education and Macmillan.
xvi. Moyinoluwa, T. D. (2015). Implementation of the revised 9-year basic education curriculum (BEC) in the North Central Nigeria: A Monitor of Benue State. *Journal of Research and Method in Education, 5*(3), 67-72
xvii. Muraki, (1992). *A* generalized partial credit model: Application of an Em Algorithm. *Applied Psychological Measurement, 16*(2), 159-176.
xviii. Nworgu, B.G (2015). *Educational research: Basic issues and methodology* (3rd Ed). Nsukka: University Trust Publisher.

xix. Nworgu, B.G. (1985). *The development and preliminary validation of physics achievement test (PAT)*. MSc Dissertation (unpublished). University of Nigeria.

xx. Nworgu B.G. & Agah J.J. (2012). Application of three parameter logistic model in the Calibration of mathematics Achievement Test. *Journal of Educational Assessment in Africa 29,* (7), 162 – 172.

xxi. Obinne, A.D.E. (2011). Psychometric analysis of two major examinations in Nigeria: Standard Error of Measurement. *International Journal of Educational Sciences,* 3(2), 137-144. Also available at – http://www.Krepublisher.com.102-journals/IJES/11JES.03000-11-web/IJES. Retrieved on 1/6/2016.

xxii. Obinne, A.E. (2008). *Comparison of psychometric properties of WAEC and NECO test item under item response theory.* Unpublished Ph.D Thesis, University of Nigeria.

xxiii. Obioma, G. (1986). *Development and validation of a diagnostic mathematics achievement test for Nigeria secondary school students.* Unpublished Ph.D Thesis. University of Nigeria.

xxiv. Odusoro, U.I. (2002). *The relative effect of computer and text-assisted programmed instruction on students' learning outcomes in Mathematics.* (Unpublished Ph.D. thesis), University of Ibadan

xxv. Suskie, L. (2004). *Assessing student learning: A common sense guide*. San Francisco, CA: Jossey-Bass.

xxvi. Tang, R.L. (1996). Test of English as foreign language monograph series*,* Polytomous Item Response Theory Models and their applications in large scale testing programme. Education Testing Service, Princeton, N ew Jersey.

xxvii. Tata, U., Abba, A., & Abdullahi, M. S. (2014). The causes of poor performance in Mathematics among public senior secondary school students in Azare Metropolis of Bauchi State, Nigeria

xxviii. Ugodulunwa, C.A. & Mutsapha A.Y. (2011*):* Using differential item functioning. Analysis for improving quality of state wide examination in Nigeria. *Journal of Educational Assessment in Africa, 28*(5), 241 – 252.

xxix. Uloko, E. S. & Usman, K. O. (2007). Effect of ethno-mathematics teaching approach on students' achievement and interest. *University of Agricultural Journal of Research in Science and Media, 1*(1), 81 - 91.